Statistica Sinica **13**(2003), 163-177

THE GEOMETRY OF TWO-STAGE TESTS

Michael A. Proschan

National Heart, Lung and Blood Institute, NIH

Abstract: Sample size calculations require specification of the treatment effect, but what if this is not known? Two-stage tests use the first stage to estimate the treatment effect and modify the sample size accordingly. The purpose of this paper is to unify the theory of two-stage testing based on treatment effect. The conditional error function approach of Proschan and Hunsberger (1995) is shown to be a useful way to evaluate the properties of any two-stage test. The connection between two-stage tests and positive quadrant tests is exploited to motivate certain conditional error functions.

Key words and phrases: Double sampling, type I error rate, p-value, positive quadrant tests.

1. Introduction

Sample size calculations are an important part of the planning of any well-designed experiment. For a t-test of two normal means, this involves the specification of the treatment effect, δ, and the standard deviation, σ. Several authors (for example, Wittes and Brittain (1990), Birkett and Day (1994), Betensky and Tierney (1997)) have considered two-stage designs whereby the first stage is used to estimate σ, and that estimate is used to determine the ultimate sample size required for a desired level of power. These methods are usually sufficient because in most settings one can specify a desired treatment effect. Sometimes, however, the current state of knowledge is very limited. In this situation one may wish to use the first stage to estimate the treatment effect upon which the ultimate sample size is based.

Two early adaptive methods based on the treatment effect were Bauer and Köhne (1994) and Proschan and Hunsberger (1995). Bauer and Köhne (1994) proposed a procedure based on Fisher's product of p-values. They noted that the first and second stage p-values p_1 and p_2 from any test statistic with a density are independent uniform random variables under the null hypothesis, even if the second stage sample size depends on first stage results. Thus, $-2\ln(p_1p_2)$ can be referred to a chi-squared distribution with 4 degrees of freedom regardless of how the second stage sample size is chosen. Bauer and Köhne's procedure is very general. Not only can it be used for any test statistic with a density, it

can be used even if one changes the test statistic after looking at the first stage results. Proschan and Hunsberger (1995) focused on sample size re-estimation for the comparison of normal means. Their procedure entailed prior specification of a conditional error (CE) function $A(z_1)$ dictating the amount of conditional type I error rate to use at the end of the study given the first stage z-score z_1. They concentrated on one particular $A(z_1)$ called the circular CE function. Both Bauer and Köhne (1994) and Proschan and Hunsberger (1995) allowed the possibility of stopping at the first stage if the evidence strongly supported either a treatment benefit or lack of benefit. Wassmer (1998) compared the Bauer and Köhne procedure to the circular CE function approach and concluded that they yield similar power. Several papers have either applied or expanded on the papers of Bauer and Köhne (1994) and Proschan and Hunsberger (1995); e.g., see Posch and Bauer (1999), Liu and Chi (2001), and Wassmer (1999). Lehmacher and Wassmer (1999) and Cui, Hung and Wang (1999) approached the problem from a completely different perspective. Nonetheless their methods, when restricted to two stages, can be shown to be equivalent to using a linear conditional error function.

This paper uses a geometric perspective to (1) characterize the class of two-stage, α-level tests, (2) demonstrate that two-stage α-level tests and CE functions are really two sides of the same coin, (3) show the connection between two-stage α-level tests and positive quadrant tests, and (4) show properties of different CE function tests.

2. Characterization of Two-Stage Tests

For simplicity, consider the one-tailed t-test comparing a treatment and control mean, $H_0 : \mu_C = \mu_T$ versus $H_1 : \mu_C < \mu_T$, assuming a common standard deviation σ. Assume for ease of presentation that σ is known (the Discussion section explains how this assumption can be eliminated). The value z_1 of the first stage z-score with n_1 observations per arm is used to determine the number $n_2 = n_2(z_1)$ of additional observations to take in each arm. First assume that $n_2(z_1) > 0$ for every z_1, so that a second stage is assured. We relax this assumption at the end of this section. Let z_2 be the z-score for data from the second stage only. The decision whether to reject the null hypothesis is based on the value (z_1, z_2) of the sufficient statistic, the first and second stage z-scores. The first goal is to characterize two-stage tests based on (z_1, z_2).

Under the null hypothesis Z_1 has a standard normal distribution, and the conditional distribution of Z_2 given $Z_1 = z_1$ is also standard normal, even if the second stage sample size depends on z_1. Because the conditional distribution of Z_2 given $Z_1 = z_1$ does not depend on z_1, Z_1 and Z_2 are independent. Thus, Z_1 and Z_2 are independent and identically distributed standard normals under

the null hypothesis in the adaptive sample size setting. Any α-level, two-stage test rejects the null hypothesis for values (z_1, z_2) in some fixed region $\mathcal{R}$ of the plane, where $\int\int_{\mathcal{R}} \phi(z_1)\phi(z_2)dz_1dz_2 = \alpha$ and ϕ denotes the standard normal density function. This important observation means that even though there is a rule $n_2(z_1)$ for determining the second stage, per-arm sample size, the rejection region $\mathcal{R}$ specifies an α-level procedure even if one decides to use a different second stage sample size, provided that the sample size actually used is a measurable function of z_1. For example, suppose one decided a priori to use the rejection region $\{(z_1, z_2) : (z_1 + z_2)/\sqrt{2} > 1.96\}$, where z_1 is the z-score after 100 observations/arm. Suppose further that the second stage sample size rule were $n_2(z_1) = 100$ if $z_1 \leq 1$ and 50 if $z_1 > 1$. Even if one decided not to follow the sample size rule, by choosing $n_2 = 200$ observations/arm, for example, the test procedure still has level $\alpha = 0.025$ if one uses the rejection region $(z_1 + z_2)/\sqrt{2} > 1.96$. The only requirement is that the sample size one would actually use for different possible values of z_1 is a measurable function of z_1. Thus we have the following

Result 1. *Any two-stage, α-level test based on the sufficient statistic (Z_1, Z_2) corresponds to a fixed rejection region $\mathcal{R}$ in the (z_1, z_2) plane that maintains level α whether or not the original second stage sample size rule $n_2(z_1)$ is followed, provided the sample size actually used is a measurable function of z_1.*

One type of two-stage test is a CE function test. Before the experiment begins, a CE function $A(z_1)$ is specified, where $0 \leq A(z_1) \leq 1$ and $\int A(z_1)\phi(z_1)dz_1 = \alpha$. Having observed $Z = z_1$, one is allowed to use a conditional type I error rate of $A(z_1)$. Operationally, one chooses n_2 additional observations/arm and rejects the null hypothesis if the z-score using all $2(n_1 + n_2)$ observations,

$$z = \frac{\sqrt{n_1}z_1 + \sqrt{n_2}z_2}{\sqrt{n_1 + n_2}}, \tag{1}$$

exceeds critical value $c = c(z_1)$,

$$c = \frac{\sqrt{n_1}z_1 + \sqrt{n_2}z_A}{\sqrt{n_1 + n_2}}, \tag{2}$$

where z_A is shorthand notation for $\Phi^{-1}\{1 - A(z_1)\}$. Proschan and Hunsberger (1995) showed that this is an α-level procedure for any n_2.

It is easy to see from (1) and (2) that rejection of the null hypothesis is equivalent to $z_2 > z_A$ (or $\geq$). Thus for fixed z_1, rejection occurs for z_2 exceeding a constant. Such a test is called an *increasing* test because the test function $\psi(z_1, z_2) = I\{(z_1, z_2) \in \mathcal{R}\}$ is increasing in z_2. Only increasing tests should be considered because the likelihood ratio $f(z_1, z_2 \mid \mu_T - \mu_C = a > 0)/f(z_1, z_2 \mid$

$\mu_T - \mu_C = 0$) is an increasing function of z_2 for fixed z_1, regardless of the sample size rule $n_2(z_1)$.

Result 2. *There is a* 1-1 *correspondence between increasing two-stage tests* $\psi(z_1, z_2)$ *and CE functions* $A(z_1)$ *as follows. With any increasing test function* $\psi(z_1, z_2)$ *associate the CE function* $A(z_1) = \int_{-\infty}^{\infty} \psi(z_1, z_2)\phi(z_2)dz_2$*; with any CE function* $A(z_1)$ *associate the increasing test function* $\psi(z_1, z_2) = I(z_2 > z_A)$*. The following rejection rules are equivalent:*

1. $\psi(z_1, z_2) = 1$.
2. $z_2 > z_{A(z_1)}$ (or $\geq$),
3. $z > c$ *(or* $\geq$*), where* z *and* c *are given by* (1) *and* (2)*, respectively.*

Proof. First we show that $A(z_1) = \int_{-\infty}^{\infty} \psi(z_1, z_2)\phi(z_2)dz_2$ is a CE function. Note that $A(z_1)$ is a version of $E\{\psi(Z_1, Z_2) \mid Z_1 = z_1\}$, so $\int A(z_1)\phi(z_1)dz_1 = E[E\{\psi(Z_1, Z_2) \mid Z_1\}] = E\{\psi(Z_1, Z_2)\} = \alpha$.

Next we show that 1 and 2 are equivalent. If $\psi(z_1, z_2)$ is an increasing test function, then $\psi(z_1, z_2) = 1$ is equivalent to $z_2 > b$ (or $\geq$) for some b that may depend on z_1. Furthermore, under H_0, $Z_2 \mid Z_1 = z_1$ has a standard normal distribution, so $1 - \Phi(b) = \Pr(Z_2 > b \mid Z_1 = z_1) = \int_{-\infty}^{\infty} \psi(z_1, z_2)\phi(z_2)dz_2 = A(z_1)$, hence $b = z_A$. Technically, this string of equalities holds except on a set of z_1 with probability 0.

The equivalence of 2 and 3 is immediate from (1) and (2), so the proof is complete.

An implication of Result 2 is that, rather than restricting the class of two-stage procedures, CE functions provide a useful way to evaluate them. By Result 2, condition 2, the two-stage test is equivalent to rejecting the null hypothesis when the second stage p-value is less than $A(z_1)$. Essentially, the first stage z-score dictates the α-level to be used for the second stage data, making immediate the formulas for conditional power, CP, and additional sample size, n_2, to achieve a given conditional power:

$$CP_\delta = 1 - \Phi\left(z_A - \sqrt{n_2/2}/\delta\right); \quad \text{For CP } 1-\beta, \quad n_2 = 2(z_A + z_\beta)^2/\delta^2, \qquad (3)$$

where $\delta = (\mu_T - \mu_C)/\sigma$. Thus, the properties of any two-stage test depend only on the induced CE function $A(z_1)$. Another way to look at this is that the set of points $\{(z_1, z_2) : z_2 = z_{A(z_1)}\}$ forms the boundary of the rejection region.

We now relex the assumption that $n_2(z_1) > 0$ for all z_1. If $n_2(z_1) = 0$, then we cannot talk about the conditional distribution of Z_2 given $Z_1 = z_1$ because Z_2 does not exist. But suppose we agree to artificially generate a standard normal deviate to call Z_2 in such a case. Now Z_1 and Z_2 are independent standard normals regardless of whether $n_2 = 0$, and Results 1 and 2 remain valid. For

example, suppose we modify the rejection region $\{(z_1, z_2) : (z_1 + z_2)/\sqrt{2} > 1.96\}$ to $\{(z_1, z_2) : z_1 > 2.79 \text{ or } 0 \leq z_1 \leq 2.79 \text{ and } (z_1 + z_2)/\sqrt{2} > 1.97\}$. This is a 0.025-level rejection region in the plane. If $z_1 > 2.79$, there is no point in proceeding to stage 2 because rejection of H_0 is assured. Likewise, $z_1 < 0$ obviates the need for a second stage because acceptance of H_0 is assured. The conditional error function associated with this procedure is 0 if $z_1 < 0$, and is 1 if $z_1 > 2.79$. We could imagine drawing a standard normal deviate to call z_2 in that case, and then the rejection region may still be expressed as $z_2 > z_A$ because z_A is either $\pm\infty$. This is simply a mathematical exercise that guarantees the veracity of the statement that Z_1 and Z_2 are independent standard normals (even if $n_2 = 0$).

3. Connection Between Two-Stage Tests and Positive Quadrant Tests

Under the alternative hypothesis, $(\mu_T - \mu_C)/\sigma = \delta > 0$, the marginal distribution of Z_1 and conditional distribution of Z_2 given Z_1 are normal with unit variances and means $(\theta_1, \theta_2) = (\delta/\sqrt{2})(\sqrt{n_1}, \sqrt{n_2})$. Imagine that the second stage sample size had been fixed in advance. Then Z_1 and Z_2 are independent, and we want to test the null hypothesis that $(\theta_1, \theta_2) = (0, 0)$ against the alternative hypothesis that $(\theta_1, \theta_2) = (\delta/\sqrt{2})(\sqrt{n_1}, \sqrt{n_2})$ for some $\delta > 0$. There is a specific direction $(\sqrt{n_1}, \sqrt{n_2})$ for the mean vector (θ_1, θ_2). The optimal test statistic is the usual fixed-sample z-statistic (1) for $n_1 + n_2$ observations/arm. Each different direction $(\sqrt{n_1}, \sqrt{n_2})$ yields a different optimal linear combination of z_1 and z_2. In the adaptive sample size setting, n_2 is not known a-priori, so the mean direction for $(\theta_1,\ \theta_2)$ is unknown. The mean direction sweeps out the positive quadrant $Q^+ = \{(\theta_1, \theta_2) : \theta_1 \geq 0,\ \theta_2 \geq 0\}$ as n_2 ranges from 0 to ∞. Thus, there is an analogy between two-stage tests and tests of the positive quadrant alternative. This analogy is not completely airtight because with two-stage tests, Z_1 and Z_2 are independent only under the null hypothesis. Still, positive quadrant tests are appealing in the two-stage setting.

An obvious candidate is the likelihood ratio test for the positive quadrant alternative. The rejection region $\{z_1 > k\} \cup \{z_2 > k\} \cup \{z_1 > 0,\ z_2 > 0,\ z_1^2 + z_2^2 > k^2\}$ is shown in Figure 1 (see Follmann (1998)). It is unlikely in practice that one would choose to continue the study if $z_1 < 0$. Thus, it is more appealing to lop off that portion of Figure 1 and recompute k to have an α-level procedure. If $z_1 > k$ or $z_1 < 0$, one stops at the first stage with rejection or acceptance of the null hypothesis, respectively. If $0 \leq z_1 \leq k$, called the *continuation region*, one proceeds to stage 2 and rejects if $z_2 > \sqrt{k^2 - z_1^2}$. The null conditional probability of this, given $Z_1 = z_1$, is $1 - \Phi(\sqrt{k^2 - z_1^2})$. Thus the conditional error function associated with the LRT, modified by eliminating $z_1 < 0$, is the "circular" CE

function (Figure 1)

$$A_{\text{cir}}(z_1) = \begin{cases} 0 & \text{if } z_1 < 0, \\ 1 - \Phi\left(\sqrt{k^2 - z_1^2}\right) & \text{if } 0 \le z_1 \le k, \\ 1 & \text{if } z_1 > k. \end{cases} \tag{4}$$

The values of k for $\alpha = 0.025$ and $\alpha = 0.05$ are 2.267 and 1.951, respectively.

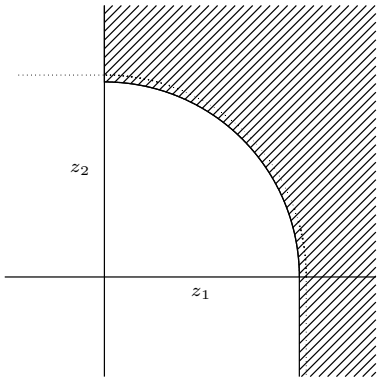

Figure 1. The rejection region for the likelihood ratio test of $\mathrm{H}_0 : (\theta_1, \theta_2) = (0, 0)$ versus the positive quadrant alternative , $Q^+ = \{(\theta_1, \theta_2) : \theta_1 \ge 0, \theta_2 \ge 0\}$ lies beyond the dotted line. Eliminating the portion with $z_1 < 0$ and decreasing the radius of the circle to maintain level α produces the circular CE function whose rejection region is shaded.

Another positive quadrant test that could be used in the two-stage setting is the likelihood ratio test for the (1,1) direction. It rejects the null hypothesis when $(z_1 + z_2)/\sqrt{2} > z_\alpha$. The null conditional probability of this, given $Z_1 = z_1$, is $1 - \Phi(\sqrt{2} z_\alpha - z_1)$. This test is a special case of the "linear" class of CE functions $z_A = a - bz_1$, $b > 0$ (Figure 2). Again one would likely not want to continue if the first stage z-score were negative, so we would lop off that portion of the rejection region.

Another modification is very desirable for a two-stage test. It would not make sense to proceed to stage 2, observe $z_2 < 0$, and then declare the treatment beneficial; if we were unconvinced of treatment benefit at stage 1 then we should be even less convinced if $z_2 < 0$. A test whose rejection region contains points (z_1, z_2) such that z_1 is in the continuation region and $z_2 < 0$ is called *inconsistent*. We can make the linear CE function tests consistent by excising the portion of the rejection region below the horizontal axis. Figure 2 shows the modified linear CE function:

$$A(z_1) = \begin{cases} 0 & \text{if } z_1 < 0, \\ 1 - \Phi(a - bz_1) & \text{if } 0 \le z_1 \le a/b, \\ 1 & \text{if } z_1 > a/b, \end{cases} \tag{5}$$

for $b = 1$ (see Section 3 for values of a that yield an α-level CE function).

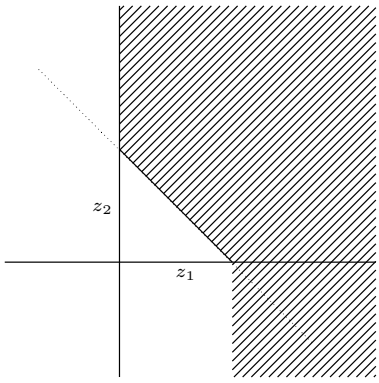

Figure 2. Linear CE function with $b = 1$ (dotted line), generated by a fixed sample procedure with interim look after planned information fraction 0.5. It corresponds to the likelihood ratio test for the $(1, 1)$ direction alternative hypothesis. In practice one might eliminate the portions with $z_1 < 0$ or points in the continuation region with $z_2 < 0$ and change the intercept slightly to maintain level α. This "modified" linear CE function has the shaded portion as its rejection region.

Bauer and Köhne (1994) used Fisher's product of independent p-values, $p_1p_2 = \{1-\Phi(z_1)\}\{1-\Phi(z_2)\}$, rejecting when $p_1p_2 < c_\alpha$, where $c_\alpha = \exp\{-\chi^2_4(\alpha)/2\}$ and $\chi^2_4(\alpha)$ is the upper α point of a chi-squared distribution with 4 degrees of freedom. This is another test that performs well when the means of Z_1 and Z_2 are both positive. Note that when $p_1 < c_\alpha$, one can stop at stage 1 because rejection of H_0 is assured. Bauer and Köhne (1994) also modified their procedure to allow early stopping for futility when the first stage p-value is less than α_0. When $\alpha_0 = 0.5$ is used, their $\alpha = 0.025$ procedure stops for benefit at stage 1 when $p_1 < 0.0102$ ($z_1 > 2.32$). If $0 \le z_1 \le 2.32$, one proceeds to stage 2, rejecting when $p_1p_2 < c_\alpha = 0.0038$. Thus, the null probability of rejecting the null hypothesis given $Z = z_1$ is $A(z_1) = \Pr(P_1P_2 \le c_\alpha \mid P_1) = c_\alpha/P_1 = c_\alpha/\{1 - \Phi(z_1)\}$.

Figure 3 shows the plot of z_A against z_1 for the circular, Bauer-Köhne, and modified linear CE functions with $b = 1$. The circular and Bauer-Köhne CE functions are very close, explaining the similar operating characteristics noted by Wassmer (1998). Even their continuation regions are nearly identical ([0, 2.27] and [0, 2.32] for the circular and Bauer-Köhne CE functions, respectively). The continuation region for the modified linear CE function is much larger ([0, 2.79]). Thus, it is more difficult to stop for benefit at stage 1 with the modified linear CE function. On the other hand, over a wide range of z_1 from about $z_1 = 0.61$ to about $z_1 = 2.18$, the value of z_2 needed to reject H_0 is smaller for the

modified linear than for the other tow CE functions (Figure 3). The explanation is simple: for fixed sample sized with $n_1 = n_2$, the linear CE function with $b = 1$ is the likelihood ratio test for the alternative that the mean of (Z_1, Z_2) is proportional to (1, 1). Thus, it does better than the other two when (z_1, z_2) is proportional to (1, 1), or at least close to it. The same is true for the modified linear CE function because it is so close to the linear CE function. On the other hand, the circular CE function is motivated by the likelihood ratio test for the entire positive quadrant, and therefore the entire spectrum of second stage sample sizes. It does reasonably well across all possible mean directions in the positive quadrant. It is not as good as the linear CE function if n_2 is close to the originally planned value n_1 (in other words, the mean vector for (Z_1, Z_2) is in the direction of (1, 1)), but is better than the linear CE function if n_2 changes substantially from what was originally planned. The Bauer-Köhne CE function is similar. The linear CE function seems preferable for trials in which it is agreed in advance that the sample size will change only modestly.

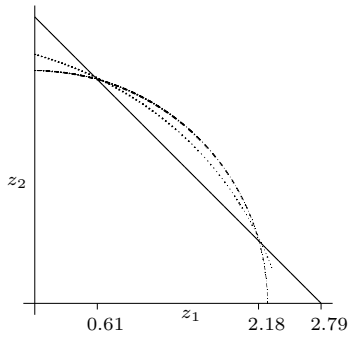

Figure 3. The circular and modified linear CE functions, along with the CE function implicit in the Bauer-Köhne (1994) procedure with $\alpha_0 = 0$. Rejection regions lie above these curves. The circular and Bauer-Köhne CE functions are quite close and have very similar continuation regions, [0, 2.27] and [0, 2.32], respectively. The modified linear CE function has continuation region [0, 2.79], which makes it more difficult to stop at the first stage. On the other hand, across a broad range of z_1 from about 0.61 to 2.18, rejection of H_0 occurs for smaller z_2 with the modified linear CE function than with the other two.

4. A New Perspective on Some Old CE Functions

4.1. A desirable but unattainable property

For a given conditional error function, the critical value c of (2) can be compared to the corresponding value for a fixed-sample z-test with the same

number of observations. It is desirable for c to be close to z_α. We would also like $c \geq z_\alpha$; it would not feel right to reject the null hypothesis using a two-stage test even though the conventional fixed sample size test statistic is not significant. Unfortunately, no CE function with continuation region $(-\infty, \infty)$ can satisfy this requirement:

Result 3. *There is no CE function with continuation region* $(-\infty, \infty)$ *such that* $c \geq z_\alpha$ *for all* z_1 *and* n_2.

Proof. Formula (2) shows that for $n_2 \to \infty$, $c \to z_A$. Requiring $c \geq z_\alpha$ for all z_1 and n_2 forces $z_A \geq z_\alpha$, which means that $A(z_1) \leq \alpha$ for all z_1. The only $A(z_1)$ such that $\int A(z_1)\phi(z_1)dz_1 = \alpha$ and $A(z_1) \leq \alpha$ for all z_1 is $A(z_1) \equiv \alpha$. Thus, the only possible candidate is the constant CE function. But if $A(z_1) \equiv \alpha$ and $z_1 = 0$, $n_2 = n_1$, then $c = z_\alpha/\sqrt{2} < c_\alpha$.

Result 3 applies when the continuation region is $(-\infty, \infty)$ but what if we accept H_0 at stage 1 if $z_1 < a$, reject if $z_1 > b$, and continue if $a \leq z_1 \leq b$? Unfortunately, this does not avoid the problem. Assuming n_1 is large, (2) shows that c is approximately z_1 when n_2 is small. Therefore, to ensure that $c \geq z_\alpha$ for all n_2, no z_1 in the continuation region can be less than z_α. But then the total type I error rate is at most $\int_{z_\alpha}^{\infty} A(z_1)\phi(z_1) \leq \alpha$, with equality if and only if $A(z_1) = 1$ for all $z_1 > z_\alpha$. In other words, the only such CE function corresponds to a fixed sample z-test on stage 1 data only.

It is simply too much to ask that $c \geq z_\alpha$ for all conceivable values of n_2, no matter how illogical they are. Choosing n_2 small when z_1 is small is bound to cause problems. When n_2 is logically tied to z_1, the critical value becomes much better behaved. A comprehensive examination of the behavior of the critical value for different CE functions is beyond the scope of this paper, but we next use a geometric perspective to prove certain desirable properties of critical values for the circular and linear CE functions.

4.2. Circular CE functions

Equations (1) and (2) show that the z-score and critical value are $\pm$ the length of the projections of (z_1, z_2) and (z_1, z_A), respectively, on $(\sqrt{n_1}, \sqrt{n_2})$. Using this geometric perspective, we can motivate the circular CE function. Suppose we used a fixed critical value c_f for the z-score (1). Proschan and Hunsberger (1995) showed that by suitable choice of $n_2(z_1)$, the type I error rate could be inflated to $1 - \Phi(c_f) + \exp(-c_f^2/2)/4$. Now consider an unrealistic, seemingly more informative scenario in which one is clairvoyant and knows at stage 1 what z_2 will be. Thus z_1 and z_2 are known, and n_2 can be chosen to maximize z. Paradoxically, incorporating this knowledge in the choice of n_2 does not further inflate the type I error rate. Figure 4 shows that when $0 \leq z_1 \leq c_f$, z is

maximized when $(\sqrt{n_1}, \sqrt{n_2})$ is in the same direction as (z_1, z_2), in which case $z = L_1 - \|(z_1, z_2)\|$. Rejection of the null hypothesis occurs for (z_1, z_2) beyond the circle of radius c_f in Figure 4. In the absence of knowledge of z_2, one can choose $(\sqrt{n_1}, \sqrt{n_2})$ in the direction of the point on the circle, $(z_1, \sqrt{c_f^2 - z_1^2})$, in which case z is the length, L_2, of the projection of (z_1, z_2) on $(z_1, \sqrt{c_f^2 - z_1^2})$. It is clear from Figure 4 that if L_1 exceeds the radius of the circle (namely c_f) then so does L_2. Thus, the null hypothesis is rejected just as often when one does not know z_2 as when he does. The rejection rates for the clairvoyant and nonclairvoyant statisticians are the same even though L_1 is always larger than L_2. The key was to choose $(\sqrt{n_1}, \sqrt{n_2})$ in the same direction as $(z_1, z_{A_{\text{cir}}})$.

Geometric reasoning can also be used to prove the following result.

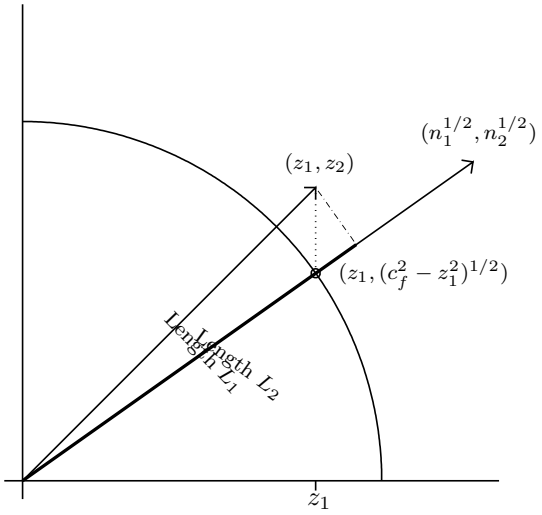

Figure 4. For two-stage tests, the usual z-score, z, for all $2(n_1 + n_2)$ observations is $\pm$ the length of the projection of (z_1, z_2) onto $(\sqrt{n_1}, \sqrt{n_2})$. Assume that $z_1 \geq 0$, $z_2 \geq 0$. If one used a fixed critical value, c_f, rejection occurs when this length exceeds c_f, meaning that the projection vector extends beyond the displayed circle of radius z_α. If one were clairvoyant and knew, at stage 1, not just z_1, but z_2 as well, one could make z as large as possible by choosing n_2 such that $(\sqrt{n_1}, \sqrt{n_2})$ is in the same direction as (z_1, z_2). Then $z = L_1$, the length of the thin line. In the absence of clairvoyance, one could choose n_2 such that $(\sqrt{n_1}, \sqrt{n_2})$ is in the same direction as $(z_1, \sqrt{c_f^2 - z_1^2})$. Then $z = L_2$, the length of the thick line. It is clear that whenever L_1 is beyond the radius of the circle, so too is L_2, even though L_1 is always larger than L_2. Thus, the maximum α-inflation is the same whether or not one is clairvoyant.

Result 4. *A consistent two-stage procedure with continuation region* $[0, k]$ (*first stage critical value* k) *is guaranteed to have second stage critical value no greater than* k *for all values of* n_2 *iff* $z_1^2 + z_A^2 \leq k^2$ *for* $0 \leq z_1 \leq k$. *Therefore, among all consistent tests with the same continuation region as* $A_{\text{cir}}(z_1)$, *the only one that guarantees that the final critical value will be no larger than that of the first stage for any* n_2 *is the test associated with* $A_{\text{cir}}(z_1)$.

Proof. It is clear that for the circular CE function, the length of the projection of (z_1, z_A) onto any vector in the positive quadrant can be no larger than the radius of the circle, which is the first stage critical value. Thus, the circular CE function has the stated property.

To see that no other CE function has this property, let $A(z_1)$ be the CE function associated with any other consistent test with continuation region $[0, k_{\text{cir}}]$. Because $\int A(z_1)\phi(z)(z)dz_1 = \int A_{\text{cir}}(z_1)\phi(z_1)dz_1$, there must be at least one z_1 for which $A(z_1) > A_{\text{cir}}(z_1)$, and hence at least one z_1 for which $\|(z_1, z_A)\| > \|(z_1, z_{A_{\text{cir}}(z_1)})\| = k$. If n_2 is selected such that $(\sqrt{n_1}, \sqrt{n_2})$ is in the same direction as (z_1, z_A), then the critical value will be $\|(z_1, z_A)\| > k$.

4.3. Linear CE functions

The simplest type of CE function is motivated by consideration of fixed sample size designs. Suppose an experiment with a fixed number, n, of patients is monitored after a fraction t of them are evaluated, $0 < t < 1$. The null conditional probability of a significant result at the end of the study, given the current z-score z_1, is

$$A(z_1) = 1 - \Phi\{(z_\alpha - t^{1/2}z_1)/(1-t)^{1/2}\}. \tag{6}$$

Now consider a different experiment with an adaptive sample size plan using the CE function (6). After observing the first stage data one may decide to continue with the originally planned sample size, in which case the critical value at the end of the study is z_α. On the other hand, one may decide to increase or decrease the final sample size, resulting in a different critical value. A CE function that is a conditional power function for a fixed sample test is said to be *generated by* that fixed sample procedure.

Note that $z_A = \Phi^{-1}\{1 - A(z_1)\} = a - bz_1$, where $a = z_\alpha/\sqrt{1-t}$ and $b = \sqrt{t/(1-t)}$. These are "linear" CE functions, in Proschan and Hunsberger's (1995) parlance. Thus, every CE function generated by a fixed-sample procedure corresponds to a linear CE function. The converse is also true; for any linear CE function with $b \geq 0$ there corresponds a fixed-sample size procedure generating it.

Result 5. *There is a* 1-1 *correspondence between the set of linear CE functions* $z_A = a - bz_1$, $b \geq 0$, *and the set of CE functions generated by fixed sample procedures with interim look at information time* t, $0 \leq t < 1$. *Specifically,* $b = \sqrt{t/(1-t)}$.

It is useful to consider two extreme case of the linear CE function. The first is $a = z_\alpha$, $b = 0$. This is generated by a fixed sample size procedure with interim look at information time 0. Regardless of the data at the first stage the amount of conditional type I error rate to use at the end of the study is α. By Result 2, this is equivalent to discarding the first stage data and rejecting the null hypothesis if $z_2 > z_\alpha$. Thus, using the CE function generated by a $t = 0$ interim look is clearly suboptimal. The second extreme case is $b \to \infty$, which is generated by a fixed sample size procedure with interim look at information time 1. In this case $A(z_1)$ tends to 0 if $z_1 < z_\alpha$ and 1 if $z_1 > z_\alpha$. In other words, this is equivalent to ignoring the second stage data and rejecting the null hypothesis if $z_1 > z_\alpha$. One should not collect any second stage data in this case.

A more reasonable procedure is exactly halfway between the two extremes, $t = 0$ and $t = 1$. Let $a = z_\alpha\sqrt{2}$, $b = 1$. This is generated by a fixed sample size procedure with interim look at $t = 1/2$. If one decided to go with the originally planned sample size, the critical value would be z_α. If one decided to enlarge or diminish the original sample size, the critical value would change somewhat, but not nearly as drastically as it would with a steep b. The fact that the critical value does not change if one uses the originally planned sample size characterizes the linear CE functions, as we see in Result 6.

Result 6. *Let* n_2^* *be a fixed number. The only CE function for which the final critical value equals* z_α *whenever* $n_2 = n_2^*$, *regardless of* z_1, *is the linear CE function generated by the fixed sample size procedure with an interim look after* n_1 *of* $n_1 + n_2^*$ *planned observations.*

Proof. As we have seen, the critical value if $n_2 = n_2^*$ is $\pm$ the length of projection of (z_1, z_2) onto $(\sqrt{n_1}, \sqrt{n_2^*})$. Figure 5 shows that for the length of this projection to exactly equal 1.96 for all z_1, (z_1, z_A) must lie on the dotted line orthogonal to $(\sqrt{n_1}, \sqrt{n_2^*})$.

As noted earlier, in practice one might modify the linear CE function to stop for futility at stage 1 when $z_1 < 0$, and eliminate $z_2 < 0$ from the rejection region. This results in the modified linear CE function specified in equation (5). Table 1 gives the intercept yielding an α-level procedure for linear and modified linear CE functions generated by a fixed sample procedure with interim look at time t. Of course other modifications are possible. One could use other continuation

regions in conjunction with the linear CE function. One could take continuation region $[-\infty, b]$, where b is any standard monitoring boundary at the first look of a two-look trial. Use of the linear CE function with this continuation region is equivalent to Lehmacher and Wassmer's (1999) and Cui, Hung and Wang's (1999) method restricted to two stages.

Table 1. Values of the intercept a and slope b for the linear and modified linear CE functions generated by a fixed sample size design with interim look after information fraction t. The first and second numbers under the a columns correspond to the linear and modified linear CE functions, respectively. With the linear CE function, there is no stopping at stage 1. One rejects H_0 at the second stage if $z_2 > a - bz_1$. With the modified linear CE function, one stops at stage 1 for futility if $z_1 < 0$ or benefit if $z_1 > a/b$, and otherwise proceeds to stage 2 and rejects H_0 if $z_2 > a - bz_1$.

t	$b = \{t/(1-t)\}^{1/2}$	a for $\alpha = 0.025$		a for $\alpha = 0.05$	
0.25	0.577	2.263	2.213	1.899	1.825
0.50	1.000	2.772	2.790	2.326	2.358
0.75	1.732	3.920	4.031	3.290	3.432

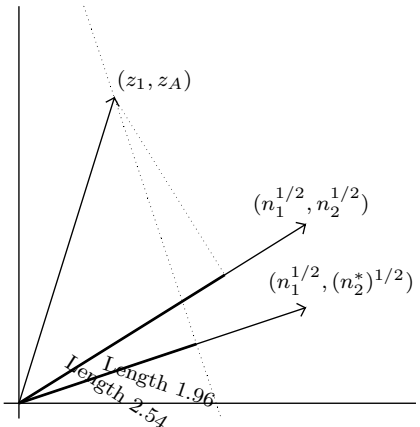

Figure 5. Projections. The critical value c for the usual z-score is $\pm$ the length of the projection of (z_1, z_A) onto $(\sqrt{n_1}, \sqrt{n_2})$. Fix n_2 at its originally planned value n_2^*. For the length of the projection to equal 1.96 irrespective of z_1, (z_1, z_A) must lie on the dotted line orthogonal to $(\sqrt{n_1}, \sqrt{n_2^*})$. In other words, z_A must be linear. This figure depicts the linear CE function $z_A = 6.20 - 3z_1$, generated by a fixed-sample procedure with $t = 0.9$. When t is so extreme, the critical value can change radically when n_2 is changed. When $z_1 = 1$ and the total sample size is increased by 25% from what was originally anticipated, c increases from 1.96 to 2.54. By contrast, if $t = 0.5$ and $z_1 = 1$, a 25% increase in sample size increases the critical value from 1.96 to only 2.005.

5. Discussion

Although for ease of presentation we assumed the standard deviation was known, Lehmacher and Wassmer (1999) have shown how to eliminate this assumption. Specifically, let $Z_i = \Phi^{-1}(1 - p_i)$, where p_i is the p-value associated with the t-statistic applied to the data from stage i, $i = 1, 2$. Then Z_1 and Z_2 are independent standard normals under the null hypothesis, just as in the case of known σ.

It is important to recognize the limitations of two-stage tests based on the treatment difference. They should not be used when sample size is based on detecting a specified minimum clinically relevant difference; then adaptively increasing the sample size risks detecting a small, clinically unmeaningful treatment effect. There is also the practical limitation that if the first-stage results are quite different than expected, the sample size may have to be drastically large than planned. That is why conditional power should be computed for several alternatives, ranging from what was originally hypothesized to what was observed in stage 1. Though they have limitations, two-stage procedures offer a way to estimate the treatment effect and project sample size using the most relevant data available, namely those from the current experiment.

References

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029-1041.

Betensky, R. A. and Tierney, C. (1997). An examination of methods for sample size recalculation during an experiment. *Statist. Medicine* **16**, 2587-2598.

Birkett, M. A. and Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statist. Medicine* **13**, 2455-2463.

Cui, L., Hung, H. M. and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853-857.

Follmann, D. A. (1998). Nonstandard multivariate tests. *Encyclopedia Statist. Sci.*, **2** (updated), 490-492.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286-1290.

Liu, Q. and Chi, Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics* **57**, 172-177.

Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical J.* **41**, 689-696.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315-1324.

Wassmer, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics* **54**, 696-705.

Wassmer, G. (1999). Mutistage adaptive test procedures based on Fisher's product criterion. *Biometrical J.* **41**, 279-293.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statist. Medicine* **9**, 65-72.

Office of Biostatistics Research, NHLBI, II Rockledge Center, 6701 Rockledge Drive, MSC 7938, Room 8222, Bethesda, Maryland 20892-7938.

E-mail: proscham@nhlbi.nih.gov

(Received June 2001; accepted June 2002)